Ethics and Information Technology
https://doi.org/10.1007/s10676-019-09516-z

ORIGINAL PAPER

Quarantining online hate speech: technical and ethical perspectives

Stefanie Ullmann[1] · **Marcus Tomalin**[2]

Abstract

In this paper we explore quarantining as a more ethical method for delimiting the spread of Hate Speech via online social media platforms. Currently, companies like Facebook, Twitter, and Google generally respond *reactively* to such material: offensive messages that have already been posted are reviewed by human moderators if complaints from users are received. The offensive posts are only *subsequently* removed if the complaints are upheld; therefore, they still cause the recipients psychological harm. In addition, this approach has frequently been criticised for delimiting freedom of expression, since it requires the service providers to elaborate and implement censorship regimes. In the last few years, an emerging generation of automatic Hate Speech detection systems has started to offer new strategies for dealing with this particular kind of offensive online material. Anticipating the future efficacy of such systems, the present article advocates an approach to online Hate Speech detection that is analogous to the quarantining of malicious computer software. If a given post is automatically classified as being harmful in a reliable manner, then it can be temporarily quarantined, and the direct recipients can receive an alert, which protects them from the harmful content in the first instance. The quarantining framework is an example of more ethical online safety technology that can be extended to the handling of Hate Speech. Crucially, it provides flexible options for obtaining a more justifiable balance between freedom of expression and appropriate censorship.

Keywords Hate speech · Social media · Ethical AI · Quarantining · Freedom of expression

Introduction

In recent years, the automatic detection of online Hate Speech (HS), and offensive language more generally, has become an active research topic in machine learning (Davidson et al. 2017; Schmidt and Wiegand 2017a, b; Fortuna and Nunes 2018). This has been prompted by increasing anxieties about the prevalence of HS on social media, and the psychological and societal harms that offensive messages can cause (Gelber and McNamara 2016; Judge and Nel 2018). In many respects, the core concerns underlying these developments are ancient ones. Harmful language (of various kinds) has been deemed problematical since at least the Old Testament, and philosophers as diverse as Aristotle, John Locke, Spinoza, Voltaire, Charles de Secondat, Baron de Montesquieu, Karl Friedrich Bahrdt, and, perhaps most famously, John Stuart Mill, have all reflected at length upon the implications of restricting free speech (for example, Aristotle 1984 [1782]; see also Britt 2010). While the specific concept of HS had already emerged by the 1940s, the social tensions in Western democracies created by increasing multi- and interculturalism from the mid twentieth century onwards led to the gradual introduction of HS-related legislation (Brown 2015, p. 182). The post-9/11 preoccupation with anti-terrorism initiatives brought a new urgency to such considerations, and recent authoritative monographs such as Ishani Maitra and Mary Kate McGowan's *Speech and Harm: Controversies over Free Speech* (Maitra and McGowan 2012), Jeremy Waldron's *The Harm in Hate Speech* (Waldron 2012), Alex Brown's *Hate Speech Law: A Philosophical Examination* (Brown 2015), and Eric Heinze's *Hate Speech and Democratic Citizenship* (Heinze 2016) have explored a wide range of practical and theoretical

✉ Stefanie Ullmann
su272@cam.ac.uk

Dr Marcus Tomalin
mt126@cam.ac.uk

[1] Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), University of Cambridge, Alison Richard Building, 7 West Road, Cambridge, Cambridgeshire CB3 9DT, UK

[2] Department of Engineering, Machine Intelligence Laboratory, University of Cambridge, Trumpington Street, Cambridge, Cambridgeshire CB2 1PZ, UK

Published online: 14 October 2019

concerns about how HS could and should be managed in liberal democracies. Some theorists, like Ronald Dworkin, have argued robustly that 'the coercive powers of the state' should be resisted even when manifest in the form of HS legislation, since free speech is a fundamental condition of legitimate government (Dworkin 2006, p. 131). Others, like Mari Matsuda or Judith Butler, believe that by not issuing legislation against HS, injurious language is, consequently, protected by the state. Specifically, Matsuda speaks of the 'victim [of HS] becom[ing] a stateless person' (Matsuda 1993, p. 25), while Butler argues that 'the state produces hate speech' in the sense that.

> […] the category cannot exist without the state's ratification, and this power of the state's judicial language to establish and maintain the domain of what will be publicly speakable suggests that the state plays much more than a limiting function in such decisions; in fact, the state actively produces the domain of publicly acceptable speech, demarcating the line between the domains of the speakable and the unspeakable, and retaining the power to make and sustain that consequential line of demarcation. (Butler 1997, pp. 76–77; emphasis in original)

In the Western world, most nations now impose penalties for some forms of expression deemed hateful because of their content, and such approaches thereby institutionalise value pluralism: the relevant legislative bodies restrict the freedoms of certain citizens so that the interests and well-being of others can be safeguarded (Galston 1999).[1] Self-evidently, these are areas where political philosophy and ethics become inextricably intertwined.

Over the last decade, though, the rapid rise of social media has created new forms of swift and efficient communication in which HS can be expressed almost instantaneously online, and often anonymously. Recognising the non-trivial problems this creates, social media providers and video-sharing platforms such as YouTube, Facebook, and Twitter have developed internal policies for HS regulation, and they also signed a Code of Conduct agreement with the European Commission (2019). At present, such decisions are taken at the corporate level, rather than the state level, which means that the companies concerned essentially regulate themselves. While there have recently been recommendations for state-level regulators, the infrastructures proposed tend to be very high-level. For instance, the UK government has outlined enforcement powers that a regulator could use against companies that failed to fulfil their duty of care (HM Government 2019, pp. 41–52). Once again, though, such methods are inherently *re*active, and this is a problem because it means that the harm has already been inflicted. In practice, the task of handling online HS involves large numbers of moderators (c 15k in the case of Facebook) checking the content of posts identified by users as being offensive. The slow and laborious nature of this inherently responsive system has motivated the conviction that automated systems are required to detect such material. And the need to define HS more precisely for these purposes (e.g., so that training data can be annotated accurately) has further emphasised the ambiguities inherent in the very phrase 'Hate Speech'. The kinds of utterances so classified vary considerably internationally, largely because HS laws take a variety of different forms depending upon the legal definitions adopted in different countries. It is also widely-recognised that the legislative focus on specific protected characteristics (e.g., race, gender, religion) has the undesirable consequence of excluding other vulnerable groups. For instance, since being an immigrant does not involve any of the protected characteristics recognised by UK law, threatening language directed towards immigrants cannot (strictly) be classified as HS, though it may still constitute a crime. This is why Facebook's public-facing 'Community Standards' document has recently been updated to indicate that the company's approach to HS overtly provides 'some protections for immigration status' (Facebook 2019a). Nonetheless, dissatisfactions with the prevailing arbitrariness and relativism of the many different definitions of HS, as well as concerns about its emphasis on the specific emotion of 'hatred', have prompted some researchers to propose alternative categories such as 'Extreme Speech' or 'Dangerous Speech' (Hare and Weinstein 2009; Benesch 2019). Nonetheless, the phrase 'Hate Speech' remains the most prominent alternative, and therefore it will be used (with reservations) in the ensuing discussion. We will interpret the phrase broadly, essentially in the same manner as Paula Fortuna and Sérgio Nunes:

> Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used. (Fortuna and Nunes 2018, p. 5)

The remaining sections of this article briefly summarise the current conventions for the regulation of online HS, before considering some of the ways in which the automatic detection of HS could be used to safeguard citizens within liberal democracies in a more ethical manner. While previous research in this area has focused primarily on the core task of developing automated methods for detecting HS, this article probes instead the way in which such technologies

[1] The most notable exception being the United States of America, which currently has no HS legislation because of the concern that it would contravene the First Amendment.

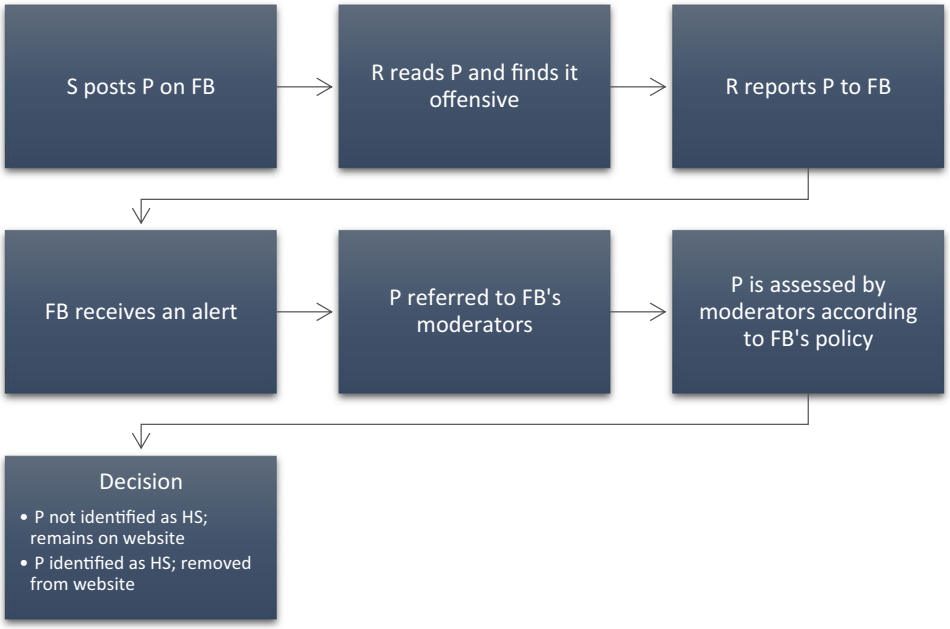

Fig. 1 Facebook's current HS regulation procedure (adapted from Allan 2017)

could be used as part of a larger infrastructure that moderates the content of social media posts in a way that does not excessively compromise freedom of expression. In particular, the method of *quarantining* is recommended as a particularly effective way of avoiding the problematical extremes of entirely unregulated free speech or coercively authoritarian censorship.

The regulation of online hate speech

As mentioned above, in response to growing public concerns about HS, most social media platforms have adopted self-imposed definitions, guidelines, and policies for dealing with this particular kind of offensive language. Continued criticism of these procedures, however, suggests that such an approach is far from ideal, therefore the current procedures adopted by Facebook, Twitter, and YouTube will be briefly summarised here as illustrative case-studies.

Figure 1 gives an overview of the way in which Facebook (FB) deals with HS if a user (S; the **S**ender) posts something (P) that another user (R; the **R**ecipient) finds offensive.

P will therefore be manually reviewed and evaluated for potential hate-inciting harmful content, if reported by a user, otherwise, it will remain visible on the website.[2] As a basis for deciding whether or not a post qualifies as HS, Facebook uses the following definition:

> We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation. (Facebook 2019b)

Additional factors are the context of the comment, cultural norms (e.g., language, country), the genre/style of the comment (e.g., humour, satire), or if a post was reproduced as a means of criticism and opposition (Allan 2017). Yet the moderation process is subject to constant revision and modification. On 28th March 2019, Facebook announced that it would block and remove white supremacist content. The decision followed heavy criticism after a live-stream of a terrorist attack in Christchurch, New Zealand, had been made available on the social media platform. New Zealand's Prime Minister Jacinda Ardern reacted saying that social media sites were 'the publisher, not just the postman' of extremist content online (BBC News 2019). Facebook's decision to take a proactive stance against the spreading of nationalist and extremist material has important consequences for HS regulation on social media more generally. Nevertheless, the company's methods for handling HS continue to be primarily *re*active rather than *pro*active.

Twitter has also been criticised repeatedly for its HS-related procedures. While its policy on 'hateful conduct' tells its users they are not permitted to 'promote violence against or directly attack or threaten other people on the

[2] Approximately 15,000 content moderators currently work for Facebook (see Newton 2019).

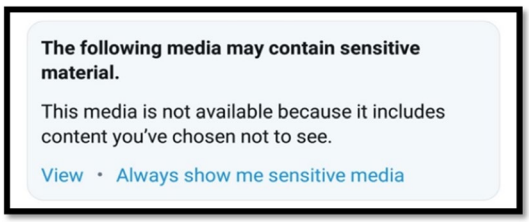

Fig. 2 Warning of potentially sensitive material on Twitter. *Source*https://twitter.com

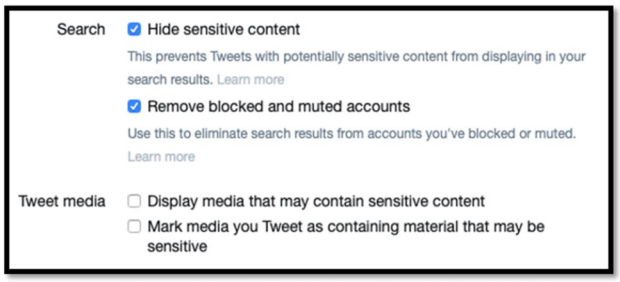

Fig. 3 Default settings for the display of potentially sensitive content on Twitter. *Source* https://twitter.com/settings/safety

basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease', critics have pointed out that this guideline is both inconsistent and ineffective, and does not protect groups from harassment (Twitter 2019; Matsakis 2018). One source of inconsistency is that Twitter, like Facebook, still relies primarily on manual and human-led HS detection and assessment. In a recent attempt to offer its users even greater protection from online harassment and abuse, the company announced a new policy prohibiting 'dehumanising speech', which it defines as:

> Language that treats others as less than human. Dehumanization can occur when others are denied of human qualities (animalistic dehumanization) or when others are denied of their human nature (mechanistic dehumanization). Examples can include comparing groups to animals and viruses (animalistic), or reducing groups to a tool for some other purpose (mechanistic). (Gadde and Harvey 2018)

The policy further states that it applies to '[a]ny group of people that can be distinguished by their shared characteristics such as their race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, serious disease, occupation, political beliefs, location, or social practices' (Gadde and Harvey 2018).

Further, Twitter has recently added an intermittent feature that blocks potentially sensitive *image*-based material (similar to Instagram's sensitive content filter), and which sends the intended recipient a warning. The recipient can then either choose to view the content, or decide to block it (see Fig. 2). In addition, users can also flag their own tweets as potentially containing sensitive material (see Fig. 3). While this feature currently functions in a somewhat arbitrary manner, the underlying approach clearly has the potential to be deployed against HS, as discussed in "The automatic detection and classification of hate speech" section. It is worth mentioning that the social news and discussion platform Reddit implemented a 'quarantine function' in 2015, which allows for content and entire communities to be put in quarantine. Content will only be released again after successful appeal (Reddit 2019), and, once again, this procedure involves careful evaluation by a human administrator.

Finally, since its emergence in 2005, YouTube has developed from being a merely video-sharing site to being an influential source of news, information, and entertainment for users worldwide. In recent years, it has provided a powerful platform for those who have sought to spread conspiracy theories (e.g. anti-vaccine), extremist views, and misinformation (see, e.g., Uscinski et al. 2018; Ottoni et al. 2018). Like Facebook and Twitter, YouTube relies on the reporting of dangerous or abusive content by already-offended users, and all such reports are subjected to qualitative review. Its 'Hate Speech Policy' (Google 2019) currently lists specific protected characteristics (11 in total, including 'veteran status'), and a reporting tool is provided that can be used to raise concerns about videos, comments, or even whole channels that promote HS.[3]

The automatic detection and classification of hate speech

Given the brisk summaries in the previous section, it should be obvious why, in the last few years, the automatic detection of HS has become an active research priority. The various systems developed so far frequently adopt a binary classification framework: given a social media post P (e.g., a tweet), the system should classify P either as constituting HS or as not constituting HS. Consequently, Precision, Recall, and Accuracy are regularly used as metrics for determining system performance. For instance, Warner and Hirshberg (2012) developed a system that classified statements as being anti-Semitic or not, while Nobata et al. (2016) and Gao et al. (2017) used the categories 'abusive' or 'clean' instead. Since the precise nature of the task and the datasets used often varies from publication to publication, numerous classification strategies have been deployed (see Fortuna and Nunes 2018 for an overview) but comparing them in

[3] The reporting tools can be found here: https://support.google.com/youtube/answer/2802027.

a meaningful manner is not always easy. While early systems tended to rely on basic word filters and simple syntactic structures to identify offensive language, more recent systems have sought to incorporate extra-linguistic knowledge-based features, as well as contextual information, to achieve more accurate detection and better classification rates. Even sociolinguistic features concerning the user's background, posting history, and online characteristics have been included in the development and training of classifiers (e.g., Dadvar et al. 2013; Schmidt and Wiegand 2017a). To consider just a few examples in greater detail, Davidson et al. (2017) used logistic regression with L1 regularization to reduce the dimensionality of the data, and they favoured a ternary classification framework in which each tweet was identified as constituting offensive language or not, with all offensive tweets subsequently being classified as constituting HS or not. Their dataset of 25 k tweets had been manually labelled (via CrowdFlower), and using Accuracy as a scoring metric, they found that 91% of the offensive tweets were being correctly identified, but only 60% of the HS (i.e., only slightly better than random chance). By contrast, Qian et al. (2018) used a Conditional Variational Autoencoder (CVAE) to distinguish among 40 hate groups with 13 different hate group ideologies, using a dataset of 3.5 million tweets. Consequently, each tweet was associated with a specific hate category label (e.g., 'ACT for America') and a specific hate speech label (e.g., white nationalist, anti-immigration). This fine-grained approach enables more specific sub-classifications of HS posts, but it depends on there being enough data associated with each sub-type. In recent years, the application of Convolutional Neural Networks (CNNs) has produced higher Precision results in many HS-related tasks (e.g., Gambäck and Sidkar 2017).

Nonetheless, despite the many technical advances, there remain persistent difficulties concerning the annotation of HS-related training data. As noted above, different researchers focus on different tasks (e.g., anti-Semitic language [i.e., a specific subset of HS], abusive language [i.e., a superset of HS]). Therefore, they use different datasets, and it is often very difficult to compare and contrast the performance of the various systems. This lack of commonly-accepted training and test datasets has significantly hampered system development. There are also problems when it comes to labelling the data. The task of classifying millions of offensive tweets is usually crowd sourced (for practical reasons), yet it is hard to guarantee quality control using that method. The subjectivity of annotators remains problematical, and it arises from diverging perceptions of what constitutes HS. This divergence is a factor even when particular definitions of HS are specified, since the perceived tone and style of a given social media post (e.g., humorous, satirical) can vary greatly. Further, there has been a growing interest in the manner in which users respond to HS. The various strategies deployed are often grouped together as instances of 'counter speech'. Crucially, users have been observed to use counter speech of different kinds, including pointing out and correcting misinformation, misrepresentations, and contradictions, warning of consequences, denouncing hateful speech, debunking via humour or sarcasm, deploying a notably positive tone, or using hostile language (Mathew et al. 2018). The important role of counter speech in countering HS is only starting to be understood, yet it clearly influences the spread of online hatred and misinformation. Clearly, more research concerning this topic (and specifically large-scale studies of datasets and the development of improved classification models) is needed.

The preceding paragraphs have offered a succinct overview of some of the recent developments in the task of detecting and classifying HS automatically. As mentioned earlier, though, the *ex post facto* identification of HS does not undo the harm that such material has already caused when posted online. While effective counter speech can certainly contribute to decreasing that harm (Butler 1997, p. 14), it would be far better to intercept potentially offensive posts at an earlier stage of the process, ideally before they have been read by the intended recipient. With this in mind, the following section will outline a framework for the automated quarantining of HS which extends an automated approach that has been used for several decades to decrease the damage caused by malware.

Quarantining hate speech

As summarised in "The regulation of online hate speech" section, social media o rganisations such as Facebook, Twitter, and YouTube currently use teams of moderators to determine whether potentially harmful posts should be removed or not. The current systems rely on already-offended users complaining about offensive messages, and the content of these is then assessed by teams of people who determine whether or not they should be deleted. These approaches will be referred to here as Too Little Too Late (TL^2) methods, since they come into effect only after the intended harm has already been inflicted, both on the direct recipient of the message, and on any indirect recipients (including the thousands of human moderators who have to encounter hundreds of examples of disturbing material every day, see Newton 2019; Simon and Bowman 2019). Consequently, TL^2 harm-reduction strategies are problematical, especially if we accept that language-mediated online harm is a serious as other sub-types (e.g., physical, financial). Also, in the influential theory of Information Ethics that Luciano Floridi (Floridi 2013, Chap. 4) has elaborated over the last few decades, there is a perceived need for an ethical framework that is primarily *patient*-oriented rather than *agent*-oriented. In other words, the moral impact of a given action is at least as

important as the decision process the relevant agent followed when electing to take that action. Viewed from this perspective, reactive TL[2] approaches are undesirable, since they do not prevent harm being caused. Inevitably, though, any proposed regulation designed to delimit harm raises familiar long-standing tensions between libertarian tendencies (e.g., freedom of expression) and more restrictive authoritarian ideologies/practices (e.g., censorship). These viewpoints have been prominent, for instance, in the important recent debates about HS legislation involving the legal theorists Dworkin (2009); Waldron (2012, 2017) and Weinstein (2017). The various disagreements have centred on topics such as whether HS bans necessarily undermine democratic legitimacy by depriving certain citizens of a voice in the political process, and diminishing their opportunity to speak without fear of criminal sanction. Contrasting views about such matters become vividly apparent in relation to online HS if the only available options are (i) to leave already-posted offensive material in situ, or (ii) to remove it entirely.

Crucially, though, these are not the only two options. Many cultures have well-established orthographic conventions for decreasing the impact of offensive written material, to differing extents. For instance, in English-speaking cultures it is often possible to replace letters in potentially derogatory words (e.g., ethnic slurs) with other symbols to render them more opaque, or else to strike out the problematical lexical items entirely. These are the sorts of content moderation methods implemented by wordfilter software in online chatrooms and forums (Roberts 2017). Here are some examples[4]:

1) You niggers are fucking retards
2) You ~~niggers~~ are ~~fucking~~ retards
3) You n*ggers are f*cking retards
4) You [redacted] are [redacted] retards

Examples (2) and (3) still enable the blatantly racist sentiment of (1) to be conveyed, but they offer different degrees and styles of textual censorship that (arguably) soften the impact of the offensive language. Option (4) is more extreme, since it entirely conceals the specific group targeted by the HS, and it merely indicates that something offensive had been written. These methods may be effective at the lexical level, but more sophisticated handling is required when HS is expressed in phrases or sentences that do not merely contain offensive words[5]:

(5) Yo if my son comes home & try's 2 play with my daughters doll house I'm going 2 break it over his head & say n my voice 'stop that's gay'.

Simple key-word spotting would not be sufficient to identify the homophobic content of (5). The term 'gay' is not inherently offensive since it is regularly used in non-pejorative ways such as 'Gay Pride' (unlike, for instance, the word 'faggot' when deployed as a homophobic slur). The import of the above tweet is only apparent at a clausal level, but even then the task of determining the referent of the demonstrative 'that' requires sentential-level comprehension. If the strategies in (2) above were to be deployed in this case, then the tweet could be modified as follows—but no current wordfiltering software could cope easily with such complex syntactic structures:

(6) Yo if my son comes home & try's 2 play with my daughters doll house ~~I'm going 2 break it over his head & say n my voice 'stop that's gay~~.

Examples like this highlight the subtleties involved in identifying and HS-related content and implementing forms of textual censorship (e.g., ellipses, strikethroughs) that ameliorate the impact of the problematical content. Given the complexity of the task, it is important to consider an alternative form of (potentially temporary) censorship—namely, quarantining. This approach has been commonplace in cyber security applications since the late 1980s, especially as a form of protection against malware.[6] For instance, Exchange Online Protection (EOP) is a spam and malware filter available as part of the Exchange Online email security service owned by Microsoft (Kjierland and Baumgartner 2018). It can be set to assess whether email is spam via EOP's own Spam Confidence Level ruleset and the detection scores assigned by the relevant email server. Any mail message that is ranked at a value of (say) five or above from either of these checks is sent to a central quarantine area, where it is retained for 15 days before being deleted. This is just one example of how quarantining is regularly deployed to protect users against software specifically designed and intended to cause particular forms of online harm (e.g., data loss, data theft, server failure).

Via analogy, HS can be viewed as another form of intentional online harm, and therefore it too can be handled by means of quarantining. The analogy is less tenuous than it

[4] This example is taken from/pol/: https://archive.4plebs.org/pol/thread/166140233 (April 2018).

[5] This example is a tweet by the comedian Kevin Hart (https://variety.com/2018/film/news/kevin-hart-responds-homophobic-tweets-1203083215/). The controversy surrounded these tweets caused him to resign as the Oscars host in 2019.

[6] Elementary quarantining methods were developed in the aftermath of the Morris worm in 1988 (see Nazario 2004, pp. 39–40).

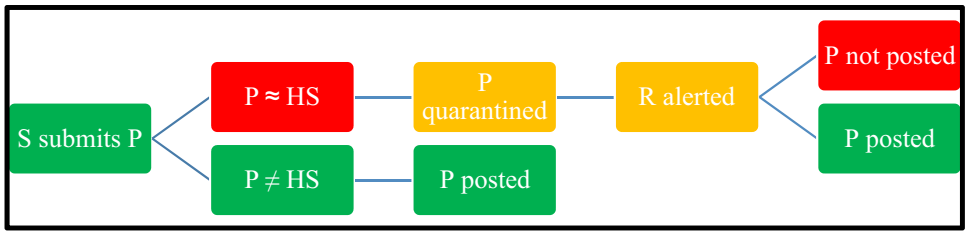

Fig. 4 The quarantining process for HS

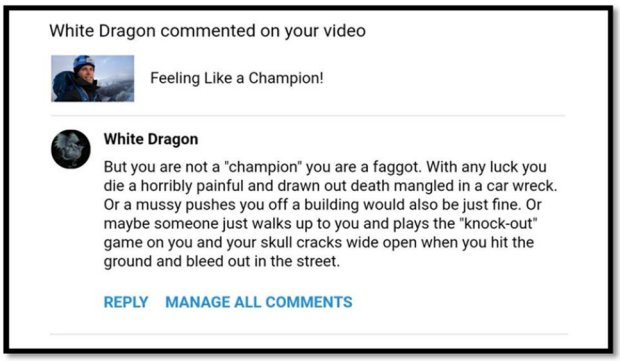

Fig. 5 Example of homophobic HS directed at American skier, Gus Kenworthy, on YouTube. *Source* https://www.youtube.com

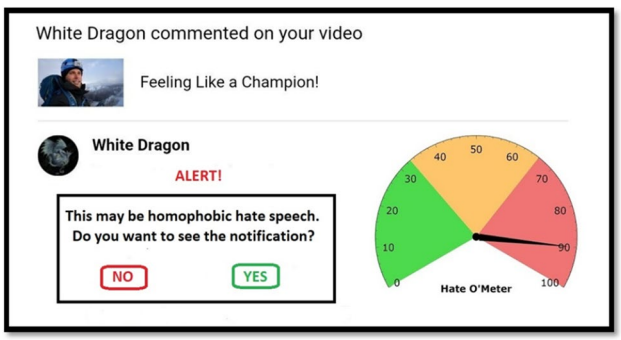

Fig. 6 Homophobic HS shown in Fig. 5 quarantined and provided with a graphic indicating degree of severity of the post

may sound initially, since online HS is sometimes generated by software (e.g., by Twitter bots) (Fig. 4). Therefore, when considered in relation to the cyber security cases, the quarantining of online HS can be viewed as an extension of existing frameworks for anti-malware protection (Daniel et al. 2019). When applied to the HS problem, quarantining is situated between the two ethical extremes of entirely permitting or entirely prohibiting the posting of certain messages. It enables the recipients of those messages (or other appropriate moderators) to decide (i) whether they wish to read the messages or not, and (ii) if they decide to read them, whether they wish to allow them to be posted or not. If this approach were adopted for a HS posting involving S (the Sender), P (the Posted message), and R (the Recipient), then the main steps in the process would be as follows:

It is helpful to consider a concrete example. The American skier, Gus Kenworthy, came out as gay in October 2015 (Fig. 5). His YouTube channel was subsequently bombarded with homophobic slurs. In 2018 he posted a video on his YouTube channel with the caption 'Feeling Like a Champion', and it received many responses including the following:

As discussed in "The regulation of online hate speech" section, there are currently very few *immediate* constraints on what anonymous users can post (Fig. 6). Consequently, the damage caused by HS of this kind is potentially instantaneous: it appears as soon as S sends it, and R can potentially read it immediately. There is no mechanism by which R can decide whether or not to read the post. There is no buffer zone. As Charles R. Lawrence emphasised as long ago as Lawrence (1990), '[t]he experience of being called "nigger," "spic," "Jap," or "kike" is like receiving a slap in the face. The injury is instantaneous' (p. 452).

By contrast, if quarantining were deployed in these cases, and if a given P were automatically identified as constituting HS, then R would receive an alert such as the following:

R could then decide whether or not to read the post after seeing who has written it (e.g., 'White Dragon') and after being informed that it has been specifically flagged up as potentially constituting homophobic HS. R could also receive an indication of the degree of severity of the post by the value specified on the Hate O'Meter graphic. This value can easily be generated from the confidence scores (continuous values in the interval [0,1]) produced by the automated HS detection system. '0' means that the post is not harmful in any way, '1' means that the post is extremely harmful/offensive.

As the above summary indicates, quarantining processes function in the intermediary ethical space between more extreme libertarian and authoritarian approaches. During quarantining, potentially offensive social media posts are neither entirely permitted nor entirely prohibited. Instead, they are held in limbo for a finite period until they have been appropriately assessed by the relevant recipients or moderators. A framework of this kind offers a better balance between positive and negative liberty (to use Isiah Berlin's well-worn distinction) than current conventions (Berlin

1969). While senders are still free to write whatever they wish (positive liberty), the recipients have the opportunity to decide which kinds of messages they wish to receive (negative liberty). The framework also offers various options concerning the degree to which this safeguarding is applied. The following implementation methods are just three of the numerous possibilities, and, as before, each scenario involve S posting the message P of which R is the direct recipient:

- The Bipartite Method: S and R (and no one else) receive an HS alert; if they both consent to the message appearing, then P becomes visible for all other indirect recipients to read.
- The Multipartite Method: even if S and R consent to post P, *all* other users (the indirect recipients) receive the alert when they first encounter P, and they can only access the contents if they give their consent.
- The Elective Method: all users can specify the degree of online HS protection they desire. For instance, in a settings file they can specify that they want to be safeguarded from, say, racist HS, and/or homophobic HS, and/or sexist HS, and so on—or simply from all kinds of HS. Consequently, users will receive alerts for any P that falls into one of the HS subtypes from which they have chosen to be protected.

Clearly, these options involve different degrees of 'friction', where, in IT-related discourse, 'friction' denotes any process that prevents users/customers accessing as rapidly as possible the goods or services they require. For instance, having to select answers from pop-up windows, having to fill in sign-up forms, failing to find relevant product specification information, and encountering long load times—these are all examples of online friction. Such experiences may annoy users, and, in extreme cases, cause them to change to other service providers (Facebook 2019c). Consequently, for many AICT developers, zero-friction systems are self-evidently an ideal. However, it should be clear from the preceding discussion that there are numerous situations in which *some* friction is highly desirable. Although social media platforms have generally tried to facilitate low-friction interactions (e.g., making it as quick and easy as possible to upload and/or share photos, audio files, documents, messages), there are situations in which this can be problematical. Adopting a high-level perspective, the Bipartite Method has the least total friction since only two users encounter the HS-related alert, while The Multipartite Method has the greatest total friction since all users encounter the alert. Crucially, in the case of The Elective Method, the degree of personal friction is chosen by each user individually. In essence they become 'voluntary consumers' of it, and the users freely opt for a degree of friction (Sumner et al. 2011, p. 18). This does not prevent them subsequently asking the service provider to remove the HS content (as is currently the case), but it does give them more control over whether or not they access that potentially harmful content in the first place. In all of these cases, the friction itself could function as a deterrent, since it is more tiresome for S to post HS if doing so constantly triggers an alert that S subsequently has to process. The same procedure could also apply to the retweeting and automatic forwarding of potentially harmful messages, since the content of those messages could also be detected automatically and an alert prompted.

These matters are of some importance because, inevitably, there are so many different scenarios in which quarantining methods could be deployed. Before considering some of these in more detail, we will distinguish formally between a Direct Recipient (DR) of a post (i.e., a person to whom the message is purposefully sent) and an Indirect Recipient (IR) (i.e., a person who might come across the post on a social media feed even though he or she was not a DR). In some situations, there is only one S and one DR (e.g., a 1-to-1 Facebook Messenger post), but in other situations there can be multiple DRs and IRs. Given this, consider the two following cases:

- Case 1: S is a neo-Nazi. S seeks to post anti-Semitic HS on his own public social media feed (i.e., S = DR). S receives an alert and consents to the message appearing; therefore, the message is posted and can be viewed by IRs who happen to read S's feed.
- Case 2: S and DR are both neo-Nazis. S seeks to post anti-Semitic HS on DR's public social media feed. S and DR both receive an alert and consent to the post; therefore, the message is posted and can be viewed by IRs who happen to read DR's feed.

In these cases, the S and the DR are both wish to propagate HS, therefore they always give their consent when the intended post triggers an alert. The concerns potentially arise when the IRs are considered. Even though an IR voluntarily chooses to view someone else's social media feed, that person may still be offended by a certain post encountered there. This is why Twitter has introduced a warning concerning 'sensitive content' (see Fig. 7). If a user's feed is already known to potentially include such material (even though it is not entirely clear how 'sensitive' should be defined), the entire page may be blocked initially, and it will only be released if specifically requested by the viewer.

How a quarantining system handles these cases would depend on the implementation method adopted. If either the Multipartite or Elective Methods were adopted, then IRs who had chosen to be fully protected from HS would receive an alert, and could then decide whether or not to view the posted message. In practice, the system could be implemented in a similar manner as Parental Guidance

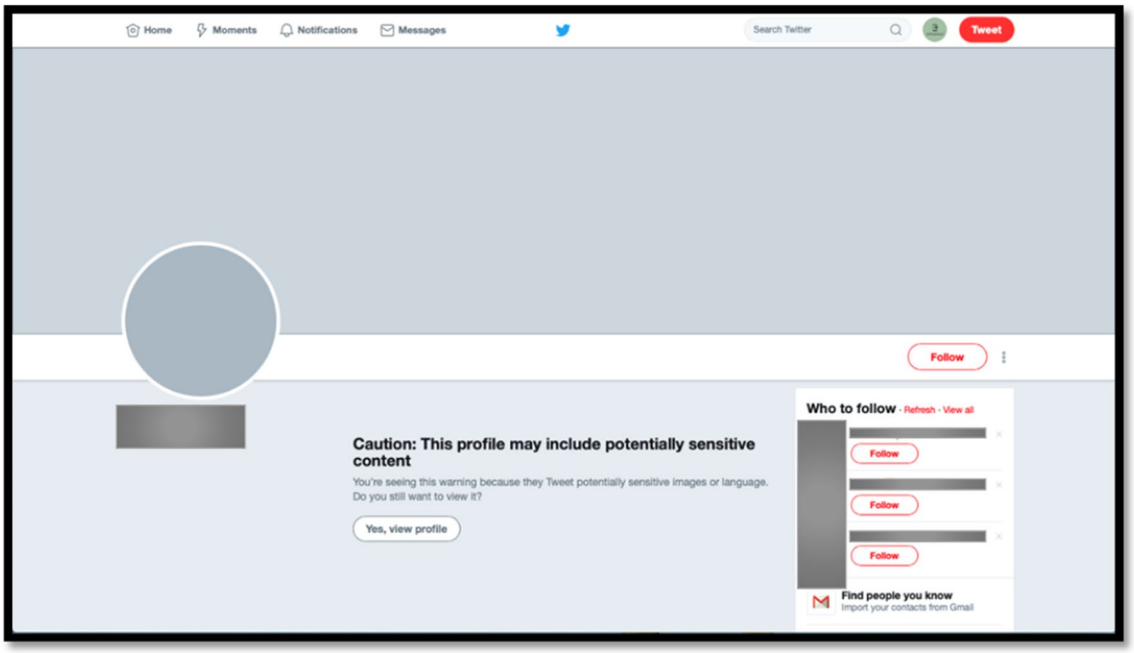

Fig. 7 Twitter warning and temporary blocking of user's feed (names and images redacted by authors. *Source* https://twitter.com

Locks on internet streaming/catch-up services such as BBC iPlayer (2019). In other words, parents could set up a social media account for their child and specify appropriate HS protection levels. If an alert were triggered, then the parents could enter a password to access and assess the quarantined post, to determine whether or not it should be posted on their child's social media feed.

While the emphasis so far in this discussion has fallen exclusively on HS, it is important to recognise that extremist groups often deploy a wide range of linguistic strategies when seeking to attract and recruit followers who may be sympathetic to their ideologies—and HS is only a part of this. Indeed, more subtle and complex techniques may not display explicit HS properties at all, the use of positive, euphemistic, and/or more abstract rhetoric may appeal to potential adherents just as effectively as HS. For instance, the terrorist group ISIS frequently used triumphant terminology like 'brothers rise up' and 'claim victory' in their recruitment strategies on social media (see, e.g., Awan 2017), while the controversial British far-right activist Tommy Robinson has repeatedly claimed he is attacking the 'fascist ideology' of Islam rather than Muslims specifically (Union Magazine 2015). Clearly, the recruitment functions of such discourses need to be examined attentively, and their diverse and multifaceted nature goes far beyond the specific task of HS detection for the purposes of quarantining.

Conclusion

This article has explored the problem of online text-based HS, and the ethical implications of the various strategies for dealing with this problematical phenomenon have been discussed. The current TL^2 regulatory frameworks were described, and some of the problems resulting from this kind of reactive self-regulation were outlined. Crucially, it was suggested that they are undesirably ineffectual, especially when viewed from a patient-oriented ethical perspective. State-of-the-art methods for the automatic detection and classification of HS were then summarised, before the main emphasis shifted to the way in which these technologies might eventually be used when their performance has improved. In particular, quarantining has been explored as a viable approach that strikes an appropriate balance between libertarian and authoritarian tendencies. In this framework, HS is treated like a form of malware, and while the senders of the HS are not censored in a crude unilateral matter, the recipients of the HS are given the agency to determine how they wish to handle the HS they have received. This approach potentially preserves freedom of expression, but the harm caused by HS is still controlled in a safe fashion by those most directly affected.

Fig. 8 Example of multimodal HS content. *Source* https://me.me

The need to consider the various available strategies for handling online HS has never been more urgent. The UK government has recently stated explicitly that.

> By designing safer and more secure online products and services, the tech sector can equip all companies and users with better tools to tackle online harms. We want the UK to be a world-leader in the development of online safety technology and to ensure companies of all sizes have access to, and adopt, innovative solutions to improve the safety of their users. (HM Government 2019, p. 77)

Given this, the importance of technological infrastructures that can facilitate the development of safer and more ethical online products should be all too apparent. And handling online HS convincingly and effectively is simply one part of a complex whole.

It is crucial to recognise, though, that the various methods considered in this article are all entirely text based. That is, they rely on sequences of (written) words being analysed in ways that enable the detection and classification of HS to be accomplished automatically. Recognising this, it is important to acknowledge that online communication is frequently and increasingly multimodal, rather than monomodal. Multimodality Studies has emerged over the last 20 years or so, largely due to the pioneering work of theorists such as Gunther Kress, Carey Jewitt, Jeff Bezemer, and Theo van Leeuven. Therefore, it is now widely accepted that different modes (e.g., image, writing, gesture, music) have different semiotic affordances (e.g., the materials the image is made from, the syntax of written language), and that these are orchestrated by meaning-makers to form multimodal ensembles. For instance, an online advert might contain moving images, written text, background music, a spoken voice-over, and so on—and these all combine to convey the meaning of the advert.[7] One inevitable consequence of multimodal approaches to meaning making is that language is necessarily displaced from a position of centrality in the analytical framework. It loses its privileged status as the primary agent of meaning making, and becomes merely one of many possible ways of creating and communicating semantic content. This obviously creates problems for automated HS-detection systems that are exclusively text based. For example, consider the following (notorious) meme, which was posted on Facebook in 2013 (see Fig. 8). It was not removed, despite requests from users, even though images of women breastfeeding were removed (Bates 2013).

This blatantly misogynistic meme combines the modes of image and writing, and the offensive nature of the whole arises from the juxtaposition of the parts. Taken in isolation, the text is not necessarily problematical: it is not inherently sexist in-and-of itself, and, in certain contexts, it could presumably constitute benignly humorous advice about sexual health and family planning. However, when presented with this particular image, the meaning of the text changes, and the violently misogynistic, connotations become apparent. Nonetheless, the multimodal character of the whole means that no current text-based HS-detection systems would classify the meme as being an instance of HS. Despite the conspicuous nature of this problem, research programmes focused on multimodal approaches to HS detection and classification have only just started to emerge (Hosseinmardi et al. 2015; Zhong et al. 2016).[8] Clearly, there is much that remains to be accomplished.

Acknowledgements Research on this paper is funded by the Humanities and Social Change International Foundation.

References

Allan, R. (2017, June 27). Hard questions: Who should decide what is hate speech in an online global community? *Facebook Newsroom*.

[7] For an overview of this framework, see Kress (2010). See also Sindoni (2018).

[8] For a brief overview, see Schmidt and Wiegand (2017a), especially Sect. 3.8.

Retrieved January 28, 2019 from https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/.
Aristotle. (1984[1782]). Nicomachean ethics 5.2.1129[b]23. In Aristotle, [4th century BCE], vol. 2.
Awan, I. (2017). Cyber-extremism: Isis and the power of social media. *Society, 54*(2), 138–149. https://doi.org/10.1007/s12115-017-0114-0.
Bates, L. (2013, May 29). The day the Everyday Sexism Project won - and Facebook changed its image. *The Independent*. Retrieved April 12, 2019 from https://www.independent.co.uk/voices/comment/the-day-the-everyday-sexism-project-won-and-facebook-changed-its-image-8636661.html.
BBC. (2019). *What is the Parental Guidance Lock?* Retrieved April 12, 2019 from https://www.bbc.co.uk/iplayer/help/how-to-guides/parental-guidance/parental_guidance_info.
BBC News. (2019, March 28). *Facebook to ban white nationalism and separatism*. Retrieved April 12, 2019 from https://www.bbc.co.uk/news/world-us-canada-47728471.
Benesch, S. (2019). The Dangerous Speech Project. Retrieved April 12, 2019 from https://dangerousspeech.org/.
Berlin, I. (1969). Two concepts of liberty. In I. Berlin (Ed.), *Four essays on liberty* (pp. 118–172). Oxford: Oxford University Press.
Britt, B. M. (2010). Curses left and right: Hate speech in the biblical tradition. *Journal of the American Academy of Religion, 78,* 633–661.
Brown, A. (2015). *Hate speech law: A philosophical examination*. Abingdon: Routledge.
Butler, J. (1997). *Excitable speech: A politics of the performative*. Abingdon: Routledge.
Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In Proceedings of the 35th European Conference on IR Research: Advances in Information Retrieval. Moscow: European Conference on IR Research.
Daniel, F., Cappiello, C., & Benatallah, B. (2019). Bots acting like humans: Understanding and preventing harm. *IEEE Internet Computing, 23*(2), 40–49.
Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM),* . (pp. 512–515). Atlanta: ICWSM.
Dworkin, R. (2006). A new map of censorship. *Index On Censorship, 35*(1), 131–133.
Dworkin, R. (2009). Foreword. In I. Hare & J. Weinstein (Eds.), *Extreme speech and democracy*. Oxford: Oxford University Press.
European Commission. (2019, March 18). Countering illegal hate speech online #NoPlace4Hate. *European Commission Justice and Consumers Newsroom*. Retrieved April 12, 2019 from https://dangerousspeech.org.
Facebook. (2019). *Community standards Part III: Objectionable content*. Retrieved April 12, 2019 from https://www.facebook.com/communitystandards/objectionable_content.
Facebook. (2019). *Community standards: Hate speech*. Retrieved February 18, 2019 from https://www.facebook.com/communitystandards/objectionable_content.
Facebook. (2019). *Zero friction future*. Retrieved April 12, 2019 from https://www.facebook.com/business/m/zero-friction-future.
Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.
Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in Text. *ACM Computing Surveys, 51*(4), 1–30.
Gadde, V., Harvey, D. (2018). Creating new policies together. Twitter Blog, 25 September 2018. Retrieved February 18, 2019 from https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html.
Galston, W. A. (1999). Value pluralism and liberal political theory. *The American Political Science Review, 93*(4), 769–778.
Gambäck, B., & Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*. https://doi.org/10.18653/v1/W17-3013.
Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 1)
Gelber, K., & McNamara, L. J. (2016). Evidencing the harms of hate speech. *Social Identities, 22*(3), 324–341.
Google. (2019). *YouTube help: Hate speech policy*. Retrieved April 12, 2019 from https://support.google.com/youtube/answer/2801939?hl=en-GB.
HM Government. (2019). *Online Harms White Paper*, April 2019. Retrieved April 12, 2019 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.
Heinze, E. (2016). *Hate speech and democratic citizenship*. Oxford: Oxford University Press.
Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the Instagram social network. Computing Research Repository (CoRR) https://arxiv.org/abs/1503.03909.
Judge, M., & Nel, J. A. (2018). Psychology and hate speech: A critical and restorative encounter. *South African Journal of Psychology, 48*(1), 15–20.
Kjierland S, Baumgartner P (2018) Anti-spam and anti-malware protection [EOP]. Microsoft Docs, Retrieved April 15, 2019 from https://docs.microsoft.com/en-us/office365/servicedescriptions/exchange-online-protection-service-description/anti-spam-and-anti-malware-protection-eop.
Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.
Lawrence, C. R. (1990). If he hollers, let him go: Regulating hate speech on campus. *Controversies in Constitutional Law, 39*(3), 431–483.
Maitra, I., & McGowan, M. K. (2012). *Speech and harm: Controversies over free speech*. New York: Oxford University Press.
Mathew, B., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherje, A. (2018). Thou shalt not hate: Countering Online Hate Speech. ICWSM 2019.https://doi.org/10.13140/RG.2.2.31128.85765.
Matsakis, L. (2018). Twitter releases new policy on 'dehumanizing speech'. Wired, 25 September 2018. Retrieved February 18, 2019 from https://www.wired.com/story/twitter-dehumanizing-speech-policy/.
Matsuda, M. J. (1993). Public response to racist speech: Considering the victim's story. In M. J. Matsuda, C. R. Lawrence III, R. Delgado, & K. Williams (Eds.), *Words that wound: Critical race theory, assaultive speech, and the first amendment* (pp. 17–52). New York: Routledge.
Nazario, J. (2004). *Defense and detection strategies against internet worms*. Boston: Artech House.
Newton C (2019) The trauma floor: The secret lives of Facebook moderators in America. The Verge, 25 February 2019, Retrieved April 01, 2019 from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, (pp. 145–153).

Ottoni, R., Bernardina, P., Cunha, E., Meira, W., Jr., Magno, G., & Almeida, V. (2018). Analyzing right-wing YouTube channels: Hate. *Violence and Discrimination*. https://doi.org/10.1145/3201064320108 1.

Qian, J., El Sherief, M., Belding-Royer, E. M., & Wang, W. Y. (2018). Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (pp. 3550–3559).

Reddit. (2019). Account and community restrictions: Quarantining subreddits. Retrieved August 12, 2019 from https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits.

Roberts, S. T. (2017). Content moderation. In L. Schintler & C. McNeely (Eds.), *Encyclopedia of big data*. Cham: Springer. https://doi.org/10.1007/978-3-319-32001-4.

Schmidt, A., Wiegand, M. (2017). A survey on hate speech detection using natural language processing'. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10).

Schmidt, A., Wiegand, M. (2017). A survey on hate speech using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media* (pp. 1–10).

Simon, S., Bowman, E. (2019). Propaganda, hate speech, violence: The working lives of Facebook's content moderators. NPR, 02 March 2019, Retrieved April 01, 2019 from https://www.npr.org/2019/03/02/699663284/the-working-lives-of-facebooks-content-moderators.

Sindoni, M. G. (2018). Direct hate speech versus indirect fear speech: A multimodal critical discourse analysis of the Sun's editorial "1 in 5 Brit Muslims' sympathy for jihadis". *Lingue Linguaggi, 28,* 267–292.

Sumner, L. W. (2011). Criminalizing expression: Hate speech and obscenity. In J. Deigh & D. Dolinko (Eds.), *The Oxford handbook of philosophy of criminal law* (pp. 17–33). Oxford: Oxford University Press.

Twitter. (2019). *Hateful conduct policy*. Retrieved April 12, 2019 from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

Union Magazine. (2015). *Tommy Robinson interview for UNION magazine Edition#2*. Retrieved August 18, 2019 from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policyhttps://www.youtube.com/watch?v=5dRiBRM-BD8.

Uscinski, J. E., DeWitt, D., & Atkinson, M. D. (2018). A web of conspiracy? internet and conspiracy theory. In A. Dyrendal, D. G. Robertson, & E. Asprem (Eds.), *Handbook of conspiracy theory and contemporary religion* (pp. 106–130). Leiden: Brill.

Waldron, J. (2012). *The harm in hate speech*. Cambridge: Harvard University Press.

Waldron, J. (2017). The conditions of legitimacy: A response to James Weinstein. *Constitutional Commentary, 32,* 697–714.

Warner, W., Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, (pp. 19–26).

Weinstein, J. (2017). Hate speech bans, democracy, and political legitimacy. *Constitutional Commentary, 32,* 527–583.

Weinstein, J., & Hare, I. (2009). General introduction: Free speech, democracy, and the suppression of extreme speech. *Past and Present*. https://doi.org/10.1093/acprof:oso/9780199548781.003.0001.

Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., & Caragea, C. (2016). Content-driven detection of cyberbullying on the Instagram social network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, (pp. 3952–3958).